DOCUMENT RESUME

ED 407 428 TM 026 466

AUTHOR Vidal, Sherry
TITLE Regression Is a Univariate General Linear Model Subsuming Other Parametric Methods as Special Cases.
PUB DATE Jan 97
NOTE 21p.; Paper presented at the Annual Meeting of the Southwest Educational Research Association (Austin, TX, January 23-25, 1997).
PUB TYPE Reports - Evaluative (142) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Correlation; *Equations (Mathematics); *Regression (Statistics); *Scaling
IDENTIFIERS *General Linear Model; Statistical Package for the Social Sciences; *Univariate Analysis

ABSTRACT

Although the concept of the general linear model (GLM) has existed since the 1960s, other univariate analyses such as the t-test and the analysis of variance models have remained popular. The GLM produces an equation that minimizes the mean differences of independent variables as they are related to a dependent variable. From a computer printout of a regression analysis, the researcher can obtain weights that apply to each variable and then construct this equation. Certain univariate analyses require some variables to be in a nominal scale versus an interval scale and then provide limited information about the data when compared with other data analytic tools. This paper explains how regression subsumes all univariate analyses and how regression can provide the researcher with a greater understanding of the data. A heuristic data set using fictitious data for eight boys and eight girls from a reading test is used to clarify this discussion. Correlation is the link that ties together all univariate analyses because regression represents the model that acts as an umbrella to all univariate analyses. An appendix presents a Statistical Package for the Social Sciences (SPSS) program to illustrate regression as a GLM. (Contains 1 figure, 6 tables, and 15 references.) (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

TM

ED 407 428

Running head: REGRESSION IS A UNIVARIATE GENERAL LINEAR MODEL

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND
DISSEMINATE THIS MATERIAL
HAS BEEN GRANTED BY
SHERRY VIDAL
TO THE EDUCATIONAL RESOURCES
INFORMATION CENTER (ERIC)

Regression is a Univariate General Linear Model

Subsuming Other Parametric Methods as Special Cases

Sherry Vidal

Texas A&M University 77843-4425

Paper presented at the annual meeting of the Southwest Educational Research Association, Austin, TX, January, 1997

TM026946

BEST COPY AVAILABLE

Abstract

Although the concept of the General Linear Model has existed since the 1960's, other univariate analyses such as the t-test and OVA methods have remained popular over the years. Certain univariate analyses require some variables to be in a nominal scale vs. interval scale and provide limited information about the data as compared to other data analytic tools. This paper explains how regression subsumes all univariate analyses and how regression can provide the researcher with a greater understanding of the data. A heuristic data set is used to further clarify this discussion.

Regression is a Univariate General Linear Model Subsuming Other Parametric Methods as Special Cases

Over the years graduate students continue to learn statistics with a relatively limited conceptual understanding of the foundations of elementary univariate analyses. Maxwell, Camp, and Arvey (1981) emphasized that "researchers are not well acquainted with the differences among the various measures (of association) or the assumptions that underlie their use" (p. 525). Frequently, many researchers and graduate students make assertions such as "I would rather use Analysis of Variance (ANOVA) than regression in my study because it is simpler and will be able to provide me with all the information I need." Unfortunately, comments such as these are ill-informed and can result in the use of less desirable data analytic tools.

All univariate analyses such as the T-test, Pearson correlation, ANOVA, and planned contrasts are subsumed by correlational analyses. In 1968 Cohen acknowledged that ANOVA is a special case of regression; he stated that within regression analyses "lie possibilities for more relevant and therefore more powerful exploitation of research data" (p. 426). Thus, an understanding of a model which subsumes univariate analyses is not only pertinent to any researcher, but imperative if a researcher wants to maximize findings of research data.

The general linear model is a model which subsumes many univariate analyses. The general linear model (GLM) "is a linear equation which expresses a dependent (criterion) variable as a function of a weighted sum of independent (predictor) variables" (Falzer, 1974, p. 128). Simply stated, the GLM produces an equation which minimizes the mean differences of independent variables as they are related to a dependent variable. From a computer printout of a regression analysis, the researcher can obtain weights which apply to each variable and then construct this equation. Regression as a general linear model can provide the exact same

information as a T-test or ANOVA, but this type of analysis also provides other information which can be useful. In addition, the GLM allows the researcher more flexibility regarding the type of variables that can be entered (e.g. interval vs. nominally scaled variables).

The purpose of the present paper is to illustrate the foundations of the general linear model, in terms of regression, and the advantages this analytic tool provides over other commonly used univariate methods. The present paper conceptually outlines the general linear model; further computational detail can be found in Tatsuoka (1975). Although Cohen (1968) and Falzer (1974) acknowledged the importance of the general linear model in the 60's and 70's, the use of ANOVA methods remained popular because of its computational simplicity over other methods such as regression. Computational aids such as high powered computers were unavailable to many researchers until the 1980's; therefore researchers used analytical methods which were congruent with existing technology.

Today computers can easily compute complex analyses such as regression, however the shift from OVA methods to the general linear model has been gradual. During the years 1969-1978, Wilson (1980) found that 41% of journal articles in an educational research journal used OVA methods as compared with 25% during the years 1978-1987 (Elmore & Woehlke, 1988). Researchers are beginning to recognize that the general linear model

> can be used equally well in experimental or non-experimental research. It can handle continuous and categorical variables. It can handle two, three, four or more independent variables.... Finally, as we will abundantly show, multiple regression analysis can do anything that the analysis of variance does—sums of squares, mean squares, F ratios—and more. (Kerlinger & Pedhazur, 1973, p. 3)

One of the primary advantages of the general linear model is the ability to use categorical variables or intervally-scaled variables. OVA analyses require that independent variables are categorical, therefore independent variables which are do not naturally occur as categorical must be reconfigured into categories. This process often results in a misrepresentation of what the variable actual is. Imagine eating freshly baked chocolate chip cookies where each cookie gives a variety of chocolate chips. Often children become excited by the variation of chocolate chips that result in each cookie. Next, imagine a world where each batch of chocolate chip cookies resulted in a cookie either containing one chocolate chip or two chips. In such a world, children and adults would no longer be as interested in the variety that chocolate chip cookies provided. Similarly, when a researcher dichotomizes variables, variance is decreased, thus limiting our understanding of individual differences. While variation in a cookie is not similar to individual variation, this illustration represents how reducing an interval variable (multichip cookie) into a dichotomy (one chip or two chip cookie) can change the characteristics of a variable (cookie). Pedhazur (1982) stated: "categorization of attribute variables is all too frequently resorted to in the social sciences... It is possible that some of the conflicting evidence in the research literature of a given area may be attributed to the practice of categorization of continuous variables... Categorization leads to a loss of information, and consequently a less sensitive analysis" (pp. 452-453).

Conclusively, eliminating variance from intervally scaled predictor variables can lead to misleading results. Cliff (1987) stated:

> such divisions are not infallible; think of the persons near the borders. Some who should be highs are actually classified as lows, and vice versa. In addition, the "barely highs" are classified the same as the "very highs," even though they are

different. Therefore, reducing a reliable variable to a dichotomy makes the variable more unreliable, not less. (p. 130)

Furthermore, Thompson (1986) has established that ANOVA methods tend to overestimate smaller effect sizes: "OVA methods tend to reduce power against type II errors by reducing reliability levels of variables that were originally higher than nominally scaled. Statistical significant effects are theoretically possible only when variables are reliably measured" (p. 919). Conversely, regression analyses in general "did tend to provide more accurate estimates of explained variance than did the OVA analyses. The pattern was most noticeable when sample size was small" (Thompson, p. 924).

To examine specifically how regression and correlation subsume univariate analyses, a heuristic data set is provided in Table 1 for illustration. The fictitious data set for this example was taken from Daniel (1989). The two experimental conditions are represented by the variable *group* (1=control, 2=experimental). Other independent variables are *sex* (1=male, 2=female). The sample (n=16) consisted of eight girls and eight boys. A reading posttest with an interval scale from 1-100, where one represents a low score and 100 a high score, was used as the dependent variable. These data are used to determine how these variables can help determine which of two classrooms is most appropriate for students.

INSERT TABLE 1 ABOUT HERE.

Analysis of the Data Set

The data were analyzed using SPSS for WINDOWS 1995. The following analyses were implemented: a T-test, One-way ANOVA, Two-way ANOVA, Pearson correlation, planned

contrast, and regression analysis. Appendix A present the computer program used to analyze the data.

T-Test

Since a T-test is restricted to the comparison of two means, the two means of the independent variable group were examined in relation to the dependent variable posttest. The results are shown in Table 2. In this example, the researcher is attempting to understand possible differences on the posttest score between those subjects in the control group and those subjects in the experimental group.

INSERT TABLE 2 ABOUT HERE.

A t value of .67 is the statistic commonly referred to in research journals when using this type of statistical analysis. Tatsuoka (1975) illustrated how the t value is simply a function of the correlation coefficient in the following formula:

$$t = \left| r\sqrt{N-2} / \sqrt{1-r^2} \right|$$

T- test Done Using Regression

Step by step statistics from the regression output will be used to illustrate a proof of this formula for the heuristic data set. Table 3 illustrates the statistics that are a result of the regression analysis for this heuristic data set. Initially many researchers, especially graduate students, can become overwhelmed by this information, but this paper will attempt to highlight a few important areas of these given results.

INSERT TABLE 3 ABOUT HERE

First refer to the area titled correlation, notice the correlation coefficient between group and posttest is equal to - 177. If the correlation coefficient is inserted into the previously described formula, the following result is found:

$$t = \left| -.177\sqrt{16-2} / \sqrt{1-.031} \right| = .662/.9843 = .672.$$

This t value is identical to the t value reported in Table 2, thereby supporting the premise that a t-test is a function of correlational analysis. One can refer to the common formula for regression for a proof that regression analysis is also a function of the correlation coefficient (Thompson, 1992).

In addition, Table 3 reports an R^2 value of .0341, which can be interpreted as "the proportion of Y that we can explain with the predictors [independent variables]" (Thompson, 1992, p. 10). Furthermore, an adjusted R^2 of -.0377 is reported. This adjustment is an attempt to account for various biases (see Snyder & Lawson, 1993). However, conceptually a squared value should not be negative, thus this negative value may lead the researcher to infer that this predictor variable (group) is a poor predictor for this sample.

Lastly, the regression output gives the researcher information about the sum of squares and weights for the regression equation. These figures aid our understanding in how and which variables are account for the variance explained. In addition, the sum of square values can be used to calculate an effect size which is "the degree to which the phenomenon is present in the population" (Cohen, 1988, p. 12). Dividing the sum of squares of a given variable by the total

sum of squares will yield an effect size for each variable. Further specifics of the regression equation and beta weights will be discussed later.

One-Way ANOVA Analysis

A one-way ANOVA using group as the independent variable and "ptest" as the dependent variable was executed with the results reported in Table 4. Since a one-way ANOVA is conceptually identical to a t-test, an extensive discussion of how regression subsumes a one-way ANOVA will not be presented. However, a proof can demonstrate how an ANOVA analysis, specifically the F statistic, is a function of correlational analysis in the following formula:

$$F = t^2 = (r\sqrt{N-2} / \sqrt{1-r^2})^2$$

In other words, $F = .454 = t^2 = (.674)^2$.

INSERT TABLE 4 ABOUT HERE.

Two-Way ANOVA Analysis

Next, a two-way ANOVA was conducted. The two ways were sex (male/female) and group (experimental/condition). The dependent variable was the reading posttest score. Recall that ANOVA requires both independent variables to be in a nominal scale form, thus sex and group are appropriate variables. Table 5 lists the SPSS output for the two-way ANOVA for the heuristic data.

INSERT TABLE 5 ABOUT HERE.

The two-way ANOVA gives us the same information as in a one-way, but we also main effects and interaction effects between the variables sex and group. Sum of squares for each

variable are reported as well as an F statistic. Notice how the sum of squares for group has not changed from the One-way ANOVA to the Two-way ANOVA Analysis. An effect size could also be calculated with this sum of squares information. For this data the effect size for the two-way interaction would be 175.56/11,191.94 = .0156. This interaction variable provides the researcher with further information on how the independent variables interact with each other in relation to the dependent variable.

Two-Way ANOVA Using Planned-Contrast Regression

In order to recreate the interaction in regression, a new variable must be created. See Appendix A for the appropriate SPSS commands. The new variable "A1B1", represents the group-by-sex interaction. "A1" will now represent group membership and sex will be represented by "B1". Moreover, a planned contrast is used to create orthogonal comparisons. These results are reported in Table 6.

INSERT TABLE 6 ABOUT HERE.

The first half of the output furnishes basic descriptive statistics and correlations between the contrasts. Notice how the A1, group, correlation with "ptest" is equal to -.177, this is the same result as reported earlier in Table 3. Although SPSS prints out a new summary for each variable (e.g., group, sex, group by sex), only the last summary which includes all the variables entered is used in Table 6. Refer to the column with the T values. If you square these values they will equal the F statistic reported in the ANOVA, therefore demonstrating that multiple regression can compute the same statistics as an ANOVA without requiring the predictor variables to be in a nominal scale.

As stated earlier, a t-test may provide a researcher with the information that two means are different, but regression can inform the researcher more distinctly how two variables are different from one another in relation to the dependent variable. In regression the researcher can determine what parts of the dependent variable (y) are explained (y') or unexplained (error) by the independent variables. A Venn diagram in Figure 1 illustrates this concept in terms of the example presented earlier. The Y' area is a synthetic variable which describes the total area explained by the 3 variables ("A1", "B1", "A1B1").

INSERT FIGURE 1 ABOUT HERE

As you may notice, the correlation of each variable with Y equals beta^2. This result occurs only when effects are uncorrelated such as in orthogonal contrasts. The Y' area can also be referred to as R^2. An R^2 of .07069 is reported in the regression output, indicating that 7% of the variance can be explained by the predictors. The Venn diagram can help a researcher visualize more clearly this percentage of variance explained by the predictors.

Ultimately regression provides the researcher with an equation which gives the best possible prediction of Y' for the sample data. The basic linear equation for regression is

$$Y' = a + b_1(A1) + b_2(B1) + b_3(A1B1)$$

In standardized form, the regression equation would be:

$$Y' = \beta_1(Z_{A1}) + \beta_2(Z_{B1}) + \beta_3(Z_{A1B1})$$

See Thompson (1988) for a further discussion of beta coefficients and structure coefficients in terms of interpreting results of a regression equation. Since the variables group and sex are not in z score form, the appropriate equation would be the unstandardized regression equation.

$$Y' = 70.56 - 4.68(A1) + 4.06(B1) + 3.31(A1B1)$$

The beta values are given in Table 6. For this sample data, this equation can help a researcher determine the best possible prediction of Y', reading posttest scores, given the group condition (experimental/control) and sex (male/female). Hence, the researcher is able to make more informed decisions about the contribution of variables in relation to the dependent variable. While weights could be constructed from the statistics in an ANOVA analysis, regression provides this information without any further computations and does not require the researcher to dichotomize variables.

Summary

To conclude, there are many similarities across all univariate analyses. Correlation is the link that ties these analyses together because regression represents the model that acts as an umbrella to all univariate analyses. That is, all analyses are correlational, although some designs may not be.

References

Cliff, N. (1987). Analyzing multivariate data. San Diego: Harcourt Brace Jovanovich.

Cohen, J. (1968). Multiple regression as a general data-analytic system. Psychological Bulletin, 70, 426-443.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Hillside, NJ: Erlbaum.

Daniel, L. (1989, March). Commonality analysis with multivariate data sets. Paper presented at the annual meeting of the American Educational Research Association, San Fransisco. (ERIC Document Reproduction Service No. ED 314 483)

Elmore, R., & Woehlke, P. (1988). Statistical methods employed in the American Educational Research Journal, Educational Researcher, and Review of Educational Research from 1978 to 1987. Educational Researcher, 17(9), 19-20.

Falzer, P. (1974). Representative design and the general linear model. Speech Monographs, 41, 127-138.

Kerlinger, F. N., & Pedhazur, E. J. (1973). Multiple regression in behavioral research. New York: Holt, Rinehart, and Winston.

Maxwell, S., Camp, C., & Arvey, R. (1981). Measures of strength of association: A comparative examination. Journal of Applied Psychology, 66(5), 525-534.

Pedhazur, E.J. (1982). Multiple regression in behavioral research: Explanation and prediction. (2nd ed.). New York:Holt, Rinehart, and Winston.

Snyder, P., & Lawson, S. (1993). Evaluating results using corrected and uncorrected effect size estimates. The Journal of Experimental Education, 61, (4), 334-349.

Statistical Package for the Social Sciences (SPSS) [Computer Software]. (1995). Chicago IL: SPSS Inc.

Tatsuoka, M. (1975). The general linear model: A "new" trend in analysis of variance. Champaign, IL: Institute for Personality and Ability Testing.

Thompson, B. (1986). ANOVA versus regression analysis of ATI designs: An empirical investigation. Educational and Psychological Measurement, 46, 917-928.

Thompson, B. (1992, April). Interpreting regression results: Beta weights and structure coefficients are both important. Paper presented at the annual meeting of the American Educational Research Association, San Francisco. (ERIC Document Reproduction Service No. ED 344 897)

Willson, V. (1980). Research techniques in AERJ articles: 1969 to 1978. Educational Researcher, 9(6), 5-10.

Figure 1

Table 1

Heuristic Data

GROUP	PTEST	SEX	IQ	OVAIQ
1	18	1	93	1
1	84	2	88	1
2	64	1	85	1
2	81	2	95	1
1	98	1	93	1
1	55	2	95	1
2	49	1	85	1
2	14	2	87	1
1	99	1	130	2
1	84	2	117	2
2	47	1	118	2
2	99	2	106	2
1	83	1	118	2
1	81	2	112	2
2	74	1	103	2
2	99	2	104	2

Table 2

T-test SPSS Printout

Variable	N	Mean	SD	SE of Mean
GROUP 1	8	75.2500	26.768	9.464
GROUP 2	8	65.8750	28.847	10.199

Mean Difference = 9.3750

Levene's Test for Equality of Variances: F= .132 P= .722

t-test for Equality of Means

Variances	t-value	df	2-Tail Sig	SE of Diff	95% CI for Diff
Equal	.67	14	.511	13.943	(-20.466, 39.216)
Unequal	.67	13.92	.511	13.913	(-20.482, 39.232)

Table 3

Table 3

Multiple Regression SPSS Printout

Variable	Mean	Std Dev
PTEST	70.563	27.315
GROUP	1.500	.516

N = 16

	Correlation	
	Ptest	Group
PTEST	1.000	-.177
GROUP	-.177	1.000

Multiple R	.17723		
R Square	.03141	R Square Change	.03141
Adjusted R Square	-.03777	F Change	.45403
Standard Error	27.82647	Signif F Change	.5114

Analysis of Variance

Source	DF	Sum of Squares	Mean Square
Regression	1	351.56250	351.56250
Residual	14	10840.37500	774.31250

F = .45403 Signif F = .5114

Variable	B	SE B	95% CI	B	Beta
GROUP	-9.375000	13.913236	-39.215922	20.465922	-.177235
(Constant)	84.625000	21.998757	37.442360	131.807640	

Table 4

One-Way ANOVA SPSS Printout

Analysis of Variance

Source	D.F.	Sum of Squares	Mean Squares	F Ratio	F Prob.
Between Groups	1	351.5625	351.5625	.4540	.5114
Within Groups	14	10840.3750	774.3125		
Total	15	11191.9375			

Group	N	Mean	Standard Deviation	Standard Error	95% CI		
Grp 1	8	75.2500	26.7675	9.4637	52.8718	TO	97.6282
Grp 2	8	65.8750	28.8466	10.1988	41.7587	TO	89.9913
Total	16	70.5625	27.3154	6.8288	56.0072	TO	85.1178

Table 5

Two-Way ANOVA SPSS Printout

Source of Variation	SS	DF	MS	F	Sig of F
Main Effects					
Sex	264.06	1	264.06	.30	.591
Group	351.56	1	351.56	.41	.536
(combined)	615.63	2	307.81	.35	.708
2-Way Interactions					
Sex By Group	175.56	1	175.56	.20	.661
Model	791.19	3	263.73	.30	.822
Residual	10400.8	12	866.729		
Total	11191.94	15	746.13		

Table 6

Planned Comparison SPSS Printout

Variable	Mean	Std Dev	Variance
PTEST	70.563	27.315	746.129
A1	.000	1.033	1.067
B1	.000	1.033	1.067
A1B1	.000	1.033	1.067

Correlation

	PTEST	A1	B1	A1B1
PTEST	1.000	-.177	.154	.125

Variable(s) Entered on Step Number
3.. A1B1

Multiple R .26588
R Square .07069
Adjusted R Square -.16163
Standard Error 29.44026

Analysis of Variance

	DF	Sum of Squares	Mean Square
Regression	3	791.18750	263.72917
Residual	12	10400.75000	866.72917

F = .30428 Signif F = .8218

Variable	B	SE B	Beta	T	Sig T
A1	-4.687500	7.360066	-.177235	-.637	.5362
B1	4.062500	7.360066	.153603	.552	.5911
A1B1	3.312500	7.360066	.125246	.450	.6607
(Constant)	70.56250	7.360066	9.587	.0000	

Appendix A

SPSS Program to Illustrate Regression as a GLM

```
title '690p.sps'.
set blanks=-9999 undefined warn.
data list
   file = 'a:690p.dat'  fixed records=1 table
   /1 group 1 ptest 3-4 sex 6 iQ 8-10 OVAIQ 12.
missing values group ptest sex iq ovaiq (-9999).
list variables=all/cases=500/format=numbered.
execute.
subtitle  't-test subsumes regression'.
t-test groups=group(1,2)/variables=ptest.
execute.
regression variables=ptest group
  /descriptives=mean stddev corr /statistics=all /
  dependent=ptest /enter group.
execute.
subtitle  ' oneway anova subsumes regression'.
oneway ptest by group (1,2)
 /statistics=descriptives.
execute.
subtitle ' two-way anova as a special case of regression'.
anova
 ptest  By sex(1 2) group(1 2).
execute.
subtitle '5 planned comparison subsumes regression'.
compute A1=-1.
compute B1=-1.
if (group eq 2)A1=1.
if (sex eq 2)B1=1.
compute A1B1= A1 * B1.
regression variables=ptest A1 B1 A1B1/descriptives=all/
 criteria=pin(.95)pout(.999)tolerance(.00001)/
 dependent=ptest/enter A1/enter B1/enter A1B1.
execute.
subtitle 'y prime can be computed by anova or reg'.
compute yhat = (-.177 * group) + (.153*sex).
compute e= ptest-yhat.
print formats yhat e (F8.5).
list variables = yhat e/ cases=500.
```